基于大数据在当下出版中的应用分析

丁燕伟 吴夏艳
（首都师范大学出版社，北京 100048）

摘 要：随着科学技术的发展，信息的传播方式、信息的接收方式和媒体都发生了巨大的变化。由于新媒体技术在媒体中的应用越来越广泛，人们的日常生活越来越难以脱离新媒体技术。对于出版业而言，大数据贯穿出版过程的各个环节，从问题规划、国内生产资源的生产、版面生产到营销推广。因此，媒体需要尽快进行数字化转型，结合自身实际使用大数据，利用大数据推动业务流程变革和商业模式创新。基于此，本文简要分析当前出版物中大量数据的使用，以期对出版业的发展提出建议。

关键词：大数据；出版；影视传媒；新媒体技术；应用分析　　**中图分类号：**G250.7　　**文献标识码：**A
文章编号：1671-0134（2022）03-074-03　　DOI：10.19483/j.cnki.11-4653/n.2022.03.023
本文著录格式：丁燕伟，吴夏艳．基于大数据在当下出版中的应用分析 [J]. 中国传媒科技，2022（03）：74-76.

大数据（Big Data），曾经是互联网信息技术行业的关键词，如今已进入人类视野。在技术革命和产业变革的背景下，大数据无论是作为战略、工具还是资源，都具有巨大的变革力量，影响着许多行业和领域。出版业是文化和信息产业的重要组成部分，也不例外。大数据在出版过程中创造了整个循环结构，涵盖了从主题策划、内部产品制作、制作策划到营销的各个环节。大数据不仅已经渗透到出版业的整个功能领域，而且通过考察当前出版物中大数据的使用，逐渐成为出版商提高生产力、创新能力和竞争力的重要保障。

1. 大数据与大数据技术概述

过去对大数据（Big Data）并没有统一的定义，一般的定义是：大数据或称为海量数据、巨量数据和巨大资料。[1]是指数据的数量以及规模都宏大到无法被人类拦截、管理、处理和归类为可读信息的巨量数据。大数据具有四个主要特征：数据量大、数据类型多样、处理速度快、价值密度低。在大数据时代，如此多的数据如何挖掘和使用方式影响了大数据技术的使用。大数据的使用可以很容易地分为大数据挖掘、大数据处理、大数据储存与管理、大数据分析、大数据应用、大数据应用安全等，其中大数据分析最为重要。精准的数据分析是数据应用的基础，也是大数据实现本身价值、为用户带来利益的保障。从这个角度来看，关键的大数据技术包括云计算、分布式文件系统和并行计算架构。

大数据来自云计算，与此息息相关。大数据为云计算提供分析内容，云计算为大数据分析提供基础设施。由于数据量在 PB（1024TB=1PB）、EB（1024PB=1EB）甚至 ZB（1024EB）中的升级，难免会产生大量重复或无效数据，需要大量的计算能力和可扩展的采集，云技术正好完全具备这些能力。[2]

分布式文件系统可以通过计算机网络将物理资源连接到非本地节点，从而允许多个节点形成文件系统网络。并行计算允许同时进行多个计算。分布式文件系统和并行计算机数据库是大数据时代的关键技术，hadup 结构是典型的实现模型。

2. 大数据应用的发展背景

国内外媒体将 2013 年称为“大数据年”，大数据成为热门话题。在此之前，有大量的与大数据相关的研究成果出现，其中影响最大的是一本书、一份研究报告和一份政府发展规划，分别出现在三个时间线，跨越 30 多年：首次出现在 Alvin Toffler（美国未来学家）于 1980 年 3 月发表的著作《第三次浪潮》。[3] Toffler 在书中描述了他对未来的预测和愿景，首次引用了大数据，将大数据视为“第三次浪潮的多彩运动”。其次，2001 年，“大数据”的概念首次出现在 Gartner 公司的一份研究报告中，定义了大数据大容量、多样化和速度快的特点。[4] 第三，2012 年，奥巴马政府发布了一项涉及 6 个联邦机构、投资 2 亿美元的“大数据研发计划”，首次将大数据发展提升到国家战略层面。[5]

在中国，2013 年以来大数据应用的发展环境良好，大致可以分为三个阶段：

第一阶段：2013 年——探索研究证明阶段

2013 年，一些研究对大规模数据处理技术的重要性及其实施提供了理论解释。例如，2013 年 7 月，在第五届中国数字出版物展览会上发布了 2012—2013 年中国数字出版物年度报告。报告首次对数字出版物进行了大量数据分析和挖掘，是数字出版物的发展趋势之一。

第二阶段：2014 年——真正意义上的实施阶段

2014 年，我国出台大数据政策措施。工业和信息化部、国家发展改革委、科技部、财政部等政府部门支持发展

资金和发展规划。各级政府在大型信息项目、研究项目和技术传播方面也发挥着重要作用。成立中关村大数据交易产业联盟，开始在全国各区省市建立大数据库。《关于加快培育大数据产业集群推动产业转型升级的意见》的出台，为国内大规模数据交易的管控提供了标准。

第三阶段，2015 年至今——国家战略规划阶段

在国家战略发展层面推动大数据发展始于 2015 年。2015 年，国务院印发《促进大数据发展行动纲要》（以下简称《纲要》）。[6]《纲要》将数据作为国家的主要战略资源，加快大数据的引入和深化是必然要求和必然选择。

3. 大数据在当前出版过程中的实施

3.1 大数据改变出版流程格局

对传统出版流程最常见的描述是“作者—出版商—读者”。作者是主体，读者是终端，出版商是中间过程。版权和订单是绝对中心，读者只是内容的接收者。这种开放式线性结构的最大缺点是将需求与生产划分为两个不同的主体，经验成为重要依据，论证联系缺乏必要的科学性。

大数据在出版过程中最大的应用是创建一个“读者—出版商—作者—读者”的闭环结构。读者既是起点又是终点，既是内容生产参与者也是内容消费者，是新媒体时代以受众为中心理论的典型诠释，这个出版过程体现了大数据思维下的出版理念。这个概念和实践始于电子商务，其中最具代表性的是国外的亚马逊和中国的京东。在收集和分析用户行为数据（包括搜索、查看、购买、投票等）时，生产和需求的直接联系也保证了利润最大化。

3.2 基于大数据的选题规划

大数据首先有助于主题选择和策划，这在涉及出版的新媒体公司中尤为重要。这些公司大多拥有强大的网络数据，在收集、分析和使用与传统出版物相关的结构化、半结构化和非结构化数据方面具有优势。可以随时通过社交平台和电子商务平台记录用户行为，快速准确地反映特定领域的社会热点和趋势，为规划问题提出重要建议。近年来，许多畅销书的成功都是基于将大量数据纳入选题规划。在中国最具代表性的是《大卫·贝克汉姆》，由自营电子商务公司京东于 2014 年基于对 1700 万网站用户的分析而推出。[7]

3.3 基于大数据的量产

在创作内容的过程中，大量的数据是决定作者是否应该如何表达作品的内容或历史发展方向的最佳框架。电子出版公司 Colliloquy 在这方面取得了成功。 Coliloquy 使用 Amazon Kindle 创建交互式内容，允许读者使用“选择你的冒险”模型创建角色和情节。收集读者选择创建的数据，然后发送给作者修改脚本，《饥饿游戏》正是基于这种想法的实践。

3.4 基于大数据的展示与制作

排版制作主要包括内容审核、编辑、修正、排版等。近年来，基于 XML 数据处理标准的数字生产平台已经在互联网环境中得到应用。包括用于多用途远程编辑协作的多用户在线编辑平台，为作者、读者、编辑、出版商等多种角色的实时交流和编辑应用程序提供机会。另外，生产过程中的成品数据和碎片化数据可以同时存储，便于内容的跟踪和提取。在编辑过程中，数字标注工具可以根据常用的预定义修正符号对稿件进行电子标注。海量数据库确保内容可以找到自定义出版样式，自动排版，链接不同模板，创建不同版本。基于大量数据的编辑不仅提高了编辑效率，还提高了最终产品的质量。

3.5 基于大数据的精准营销

所谓精准营销，就是要“降低营销成本，提高营销效果”，把产品送到真正需要的用户手中。通过使用广泛的信息技术进行营销，出版商和媒体不仅可以深化客户数据，还可以利用社交网络等多种平台来维持个人和互动联系，并增加或提高用户忠诚度。分析社交网络用户圈子，实施有针对性的营销活动。

亚马逊在营销数据方面做得很好。亚马逊用个性化数据驱动的推荐系统取代了之前的专家推荐系统，从而促进了销售。系统通过分析消费信息（例如买书、关注书籍等）向读者推荐书籍。除了推荐的定制系统，亚马逊也进入了营销和数据传递的重要阶段。实体预定分布利用大数据技术深入分析过往消费支出、搜索历史列表、客户购买新产品的预测、产品是否准备交付给客户或在指导前靠近客户存储，客户下单时，收到货物的时间是以“小时”而不是“天”来衡量的。交付模式中的沉默可以部分提高客户忠诚度，提高亚马逊在客户中的声誉。

总的来说，大数据在出版行业的应用还处于起步阶段，有很多问题需要探索和检验。本土媒体企业最重要的是尽快进行数字化转型，结合自身实际，开发大数据应用，利用大数据推动业务流程转型和商业模式创新。

4. 大数据在当前出版模式中的应用

4.1 基于数据分析的专题出版

大数据和分析方法的使用，旨在为数据分析寻找原理、发现规则、预测应用，专题出版就是其中的一种应用。谷歌图书馆数据库收集了从公元时期—— 20 世纪至今出版的相关数据，通过分析各个学科的数据，尤其是对高频话题进行提取和分类，具有重要的商业意义。处于讨论热度最高的“Coliloquy 模式”，核心也是专题出版，Coliloquy 使用亚马逊 Kindle 数据开发者项目开发软件，收集用户数据，特别是用户反复突出和分析的内容，分析和提取话题，确定青春、浪漫和科幻的出版方向，并公开招募作家加入团队，最近又添加了犯罪和法律惊悚主题的版本小说。 90% 的读者读过这本书（通常为 2.99 美元到 7.99 美元），67% 的读者重复读过。 “Coliloquy 模式”的成功基于对已发表数据的分析主题的定位。[8]

4.2 基于数据交互的视觉出版

可视化技术最早应用于计算机领域，它利用计算机图形和图像处理技术将数据转换成图形、音频和视频或动画与机器进行交互。它是一种结合数据呈现、数据处理和决策分析的技术。阅读体验是评价当前出版物的重要指标，基于大数据交互的可视化，不仅可以更直观、更简化各种抽象复杂的知识，在很大程度上消除人们的阅读障碍，提供高效便捷的阅读，还可以实现数据在多空间的同时展示，为人们带来 3D 的阅读体验。视觉出版是出版业的最新模式，将对出版业的发展产生革命性的影响。目前，该模型适用于儿童和科技技术的出版物，出版方向为平面与定型、静态与动态相结合的方向。

4.3 交互式数据驱动出版

Web2.0 时期最大的成就之一是维基百科的诞生，这是维基百科技术在实践中最成功的应用。在同一个开放的数据平台上，用户从不同的角度解读相同的事件或观点，个人解读的需求是新媒体受众的一个关键特征。这种想法和实践催生了一种新的出版模式——交互式出版。Storybird 的"数字历史"创建服务平台就是一个很好的示范。Storybird 是一个基于视觉叙事的公共平台。提供来自世界各地的免费插图，鼓励读者选择有趣的图像并以书面形式分享，从而为原创书籍和出版物提供了大量服务。插图的个人解读是内容互动的延续，读者的灵感与体验的融合，使每个版本都极其独特且具有自己的归属。Storybird 在全球拥有超过 200 万用户，自成立以来的两年内创造了 500 万个故事。通过线上或线下出版，它提供了一种新的思维方式和新的出版方式。

4.4 基于数据共享的合纵连横出版

在传统媒体的转型发展中，数字化是方向，战略合作是方法。创建数字内容和专业数据库(尤其是海量数据)是跨社区、跨社会、跨行、跨界合作的必然趋势和要求。[9]"农业数字图书馆"采用基于共享数据的公开公共模式。农业数字图书馆是多省共同开发的出版项目。在 9 省联合平台南昌会议上，达成了 1600 多种农业图书电子版权合作协议，统一授权中原农民出版社、江苏凤凰三农出版中心，300 种农业图书资源库以端口开放形式给予支持，联合建设《农业数字图书馆》。第一个项目 2000 本农业图书，汇集 9 省中部资源，第二个项目将扩展到全国农业出版机构。这种基于数据交换的数据共享和横向发布模式，减少了参考链接的重复工作，并允许内容再生性、多样性和资源完整性，提高数据发布质量。

结语

2013 年，阿里巴巴重组了 25 个业务部门，以收购的方式获取相关产业以及行业的数据，丰富阿里巴巴强大的数据库。2014 年，在北京的一次大型信息发布会上，阿里巴巴集团创始人在演讲中宣布，人类正在从 IT 时代向 DT 时代过渡。阿里巴巴赢得了大量的数据红利，用数据获得利益是未来的关键因素。2015 年杭州云栖大会上宣布 DT 时代是新能源时代。"这一时期的主要来源不是石油，而是数据。"在 2016 年云栖大会结束时，马云重申，"未来的趋势不仅是知识驱动的，还有智能和数据驱动的。""基于互联网和海量数据技术的未来，它为人类创造了无数的想象和空间。"虽然大多数出版公司没有丰富的信息资源、先进的技术和足够的资源，但它们应该拥有大数据才能控制出版发展的思想。当然，大数据技术在出版行业的应用还处于起步阶段，还有很多问题需要探索和检验。但它的价值和重要性为出版业的发展提供了无限的想象空间。媒

参考文献

[1] 孙颖．基于大数据时代的数字出版产业发展趋势分析 [J]. 传播力研究，2019（36）：210.

[2] 蔡云璇．基于大数据的电力营销管理创新分析 [J]. 经贸实践，2019（12）：109.

[3] 谢洪明，杨浩．大数据在商业中的研究态势与前沿热点——基于科学知识图谱的文献计量分析 [J]. 科技与经济，2019（5）：42-46.

[4] 刘兴磊．基于大数据的广电网络转型的探讨及应用分析 [J]. 中国新通信，2019（18）：101-101.

[5] 杨曦．大数据时代的出版文化与编辑角色转型 [J]. 中国传媒科技，2021（1）：3.

[6] 闫城榛，宋迪．"大数据"时代或将引爆传媒发展新格局 [J]. 中国传媒科技，2012（10）：2.

[7] 李丽．浅谈大数据在出版社交媒体中的应用 [J]. 新闻研究导刊，2019（10）：193-194.

[8] 李东燕．基于大数据在图书馆管理与服务中的应用分析 [J]. 科技经济导刊，2019（4）：168.

[9] 郝阳．落实国家大数据发展行动纲要 用大数据思维服务数字出版转型升级 [J]. 科技与出版，2016（1）：23-24.

作者简介：吴夏艳（1989-），女，山东高密，硕士，编辑，研究方向：教育学；语言学。丁燕伟（1970-），女，河北石家庄，在职研究生，副编审，研究方向：中小学语文教学教材教法研究、编辑出版。

（责任编辑：张晓婧）